Speech Prosody 2016
31 May - 3 Jun 2106, Boston, USA

On the automatic comparison and cloning of native and non-native speech prosody.

Daniel Hirst

Aix*Marseille University, CNRS, LPL UMR 7309, 13100, Aix-en-Provence, France
daniel.hirst@lpl-aix.fr

Abstract

It is notoriously difficult to evaluate L2 prosody objectively, since there is little consensus as to what constitutes a correct prosody for a given utterance. This presentation describes an automatic procedure which consists in comparing a non-native speakers production with 10 instances of the same utterance, taken from the OMProDat database, and read by native speakers. The pitch and relative syllable durations of the native and non-native versions are normalised and compared and the version from the native speaker which is most closely correlated with that of the non-native speaker is chosen as a model. The normalised pitch and syllable durations of the native speakers recording can be cloned and transferred to the L2 utterance. The original and re-synthesised versions of the learners utterance can then be used to provide both visual and auditory feedback to the language learner.
Index Terms: speech prosody, automatic analysis, visual feedback, auditory feedback, database

1. Introduction

It is widely accepted that speech prosody is an extremely important factor contributing to the intelligibility and naturalness of non-native speech [1, 2]. It is, however, extremely difficult to evaluate prosody objectively, since there is little consensus as to what constitutes a *correct* prosody for a given utterance.

In [3] I showed that a number of objective metrics, derived automatically from the f_0 curves of speakers reading comparable texts in English, French and Chinese, taken from the OMProDat database described below [8], were sufficient to discriminate between the three languages with an accuracy of over 70%. The discrimination of English and French from Chinese, moreover reached 89% and rose to 93% when speakers' sex was included as a factor in the analysis. This was an interesting result since the metrics were obtained fully automatically from the recordings without requiring any linguistic annotation of the data other than an orthographic transcription.

In [4] we applied a similar technique to recordings of English by native speakers of English and by Mandarin Chinese speakers from Shanghai. The discrimination of native and non-native speakers of English reached over 78% for the 40 five sentence passages, each recorded by ten speakers for each language group, showing that there are systematic empirical and measurable differences between the native and non-native speakers' production of f_0. Once again this comparison did not require any linguistic annotation of the recording other than an orthographic transcription.

In this presentation, I describe an approach which consists in comparing a non-native speaker's production with several instances (10) of the same utterance read by native speakers. The purpose of this is to counteract the fact that there may be more than one acceptable intonation pattern for a given utterance. The pitch and relative syllable durations of the native and non-native versions are compared and the version from the native speaker which is most closely correlated with that of the non-native speaker is chosen as a model.

The pitch and syllable durations of the native speaker's recording can then be cloned, i.e. transferred to the L2 utterance, after a process of normalisation as described below. The original and re-synthesised versions of the learner's utterance can then be can then be used to provide both visual and auditory feedback to the language learner. The three recordings of the L1 speaker, the L2 speaker and the L2 speaker with the cloned prosody of the L1 speaker for the example described in this paper are available together with the corresponding TextGrids and PitchTiers at https://uk.groups.yahoo.com/neo/groups/praat-users/files/Daniel_Hirst/Speech Prosody 2016/Data as well as multimedia files on this site.

2. The OMProDat database

This database uses the texts from the Eurom1 corpus [5], which consist of 40 continuous, thematically connected passages, each of five sentences. The passages were originally composed in the 1980s and recordings were made for 11 European languages.

New recordings of these passages were made for Korean [6] and then later for English and French, as the corpus *AixOx* [7], and the passages were translated, adapted and recorded in Mandarin Chinese. New recordings were subsequently also made of both English and Mandarin [4] read by 5 male and 5 female Mandarin speakers from Shanghai, as part of an investigation of dialect-specific characteristics of Shanghai speech that may influence the prosody in L2 speech.

Unlike the original Eurom1 corpus, for which each speaker read only 10 or 15 passages, in the OMProDat recordings, for each language, 10 speakers (5 male and 5 female) read all 40 passages.

All these recordings were integrated as part of *OMProDat*, the Open Multilingual Prosody Database [8] and may be freely downloaded from the *Speech and Language Data Repository*: http://sldr.org/sldr000725.

3. Comparing the L2 production to that of native speakers

In order to evaluate the prosody of L2 speakers, I propose here an approach making use of several recordings by native speakers. The L2 utterance to be analysed is compared to that of 10 different recordings from the OMProDat database, read by 5 male and 5 female native speakers.

 doi: 10.21437/SpeechProsody.2016-213

For this preliminary study, I limited the comparison, as an illustration of the methodology, to two simple parameters: the correlation between the syllable durations of the readings of the utterance (Figure 1), and the correlation between the stylised pitch curves as generated by the Momel algorithm [9], normalised for both pitch and time (Figure 2). In future work we hope to explore a much larger number of more complex metrics.

Phoneme and word durations were obtained from an automatic alignment of the orthographic transcription with the acoustic signal using the SPPAS software [10]. The phoneme durations were grouped into syllables by means of a Praat script, applying the *maximum onset principle* within word boundaries [11], which states that a maximum number of consonants are grouped into the onset of a syllable following language specific phonotactic constraints.

The Momel algorithm makes it possible to factor out the local effect of individual speech segments (microprosodic effects) on f_0 and to model the underlying macroprosodic contour as a smooth continuous curve. In order to neutralise speaker-specific (and particularly gender-specific) effects of overall pitch or *key*, the Momel anchor points were pitch-normalised using the OMe scale [12], where $ome(f_0) = log2(\frac{f_0}{median})$. With this scale, pitch values are represented as the deviation in octaves from the speaker's median pitch. We found that, in general, a span of one octave, centred on the median pitch value for the speaker, gives a good estimate of a speaker's normal unemphatic pitch range.

The pitch contours are then re-synthesised from the sequence of normalised anchor points, which are also time-normalised using syllable boundaries as the unit of normalisation with a fixed duration for each syllable. We are currently experimenting with other units for normalisation such as the phoneme, the word, the rhythm unit and the tonal unit (cf [13]).

4. Results

For each L1 and each L2 speaker we calculated the correlation with the recordings of the L1 speakers for both the syllable durations and the normalised pitch curves. Preliminary results suggest that both the correlation of syllable durations and the correlation of the normalised f_0 curves with that of the 10 native speakers are quite good discriminators for non-native speech.

4.1. Syllable durations

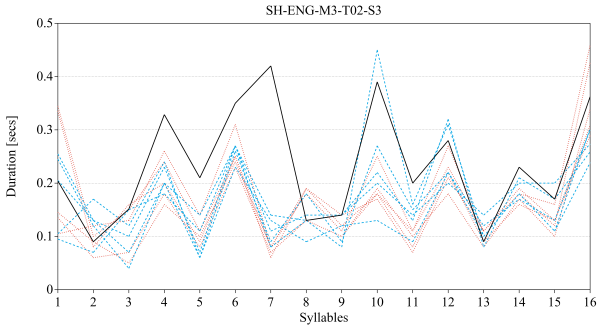
Figure (1) shows the raw syllable durations for the 10 native speakers compared to those for one male Chinese L2 speaker for the sentence T02-S3: "Could you arrange to send an engineer on Tuesday morning, please?".

For this sentence, the mean correlation for the L2 speakers ranged from 0.521 to 0.797 (excluding one atypical subject with a very low correlation) while a similar test on the recordings of the L1 speakers (also eliminating one outlier subject) ranged from 0.712 to 0.836. Taking the best correlation rather than the mean correlation gave 0.872 to 0.952 for the L1 speakers and 0.608 to 0.898 for the L2 speakers.

4.2. Pitch curves

For the f_0, the Momel anchor points were pitch-normalised using the OMe scale as described above in section 3. The anchor points were then time normalised with respect to syllable boundaries and the pitch curves were generated from the anchor-points using a fixed duration for each syllable. This made it possible to compare the pitch curves across speakers

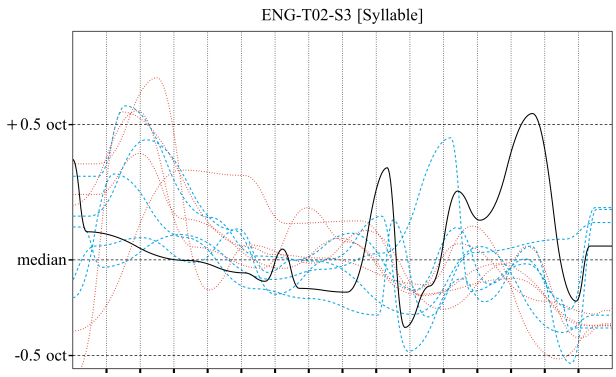
Figure 1: Raw syllable duration (secs.) for female (dotted red) and male (dashed blue) native speakers compared to that of one male Chinese L2 speaker (solid black) for the sentence T02-S3 "Could you arrange to send an engineer on Tuesday morning, please?"



without taking into account the differences of syllable durations.

Figure (2) shows the normalised pitch curves for the ten native English speakers and one male Chinese L2 speaker for the same sentence illustrated in Figure (1).

Figure 2: F0 generated from the Momel anchor points for sentence T02-S3, pitch-normalised with the OMe scale and time-normalised with respect to syllable boundaries for female (dotted red) and male (dashed blue) native speakers compared to that of one male Chinese L2 speaker (solid black).



The mean correlations between the normalised pitch curves with the L1 speakers were lower than the correlations for syllable duration, ranging from 0.176 to 0.545 for the native speakers and from 0.191 to 0.431 for the non-native speakers, once again eliminating one very untypical speaker from each group. This reflects the fact that there is far more variety in the native speakers' production of pitch than in that of syllable duration. Taking the best correlation, rather than the mean, the values ranged from 0.462 to 0.940 for the native speakers and from 0.293 to 0.677 for the L2 speakers. Just as for the syllable duration, the

best correlation for the pitch curves is a better discriminator for L1 vs L2 speakers than the mean correlation.

4.3. Both metrics

Figure (3) shows the best correlation for syllable duration vs. the best correlation for pitch for the 10 native speakers of English and the 10 Chinese L2 speakers.

Figure 3: Best correlation for syllable duration vs. best correlation for pitch for 10 native speakers of English and 10 Chinese L2 speakers, reading the sentence "Could you arrange to send an engineer on Tuesday morning please?"

It can be seen that the two correlations provide a reasonably good linear discrimination of the L1 from the L2 speakers, except for two L1 speakers (M4 and M5). Although these results are, of course, extremely preliminary, they encourage us to pursue this type of analysis in the hope that more efficient metrics can be derived automatically.

It remains, of course, to be seen whether these metrics can provide an objective evaluation of the prosody of an L2 utterance, by testing whether the metrics are correlated with the subjective evaluation of the prosody by expert listeners. This is something which we hope to address in future work.

5. Displaying the prosody of utterances

In order to provide visual feedback on the difference between the L2 speaker's prosody and that of the most similar native speaker's reading, the f_0 curves of the two recordings are modelled using the Momel algorithm and displayed using the OMe scale as described above (section 3). The duration of each syllable is represented by a circle, the diameter of which is proportional to the duration of the syllable, with the height of the circle being proportional to its mean pitch relative to the speaker's median pitch. The stylised pitch curve is then superimposed on the circles, as described in more detail in [13].

This display follows a practice common in handbooks teaching L2 intonation, see for example [14, 15], where intonation is often displayed using a representation of stressed and unstressed syllables by large and small circles respectively. Of course, it should be remembered that the diameter of the circle represents the *duration* of the syllable, and not its accentuation.

Figure (4) shows an example of the visualisation of the sentence "Could you arrange to send an engineer on Tuesday morning please?" spoken by a female native speaker of English, while Figure (5) shows the same utterance pronounced by a Chinese male speaker.

6. Cloning the prosody of a native speaker's production to an L2 utterance

In addition to providing visual feedback for an L2 speaker, the modelling technique described above could also be used to provide auditory feedback to the learner. This has not yet been tested with learners but seems to be a very promising direction to investigate.

To do this, the prosody of the L2 speaker's production is modified with Praat [16] by creating a Manipulation object. The duration of each syllable of the L2 speaker's utterance is modified by placing two points on the duration tier, the first 5 ms after the beginning of each syllable and the second 5 ms before the end. The value of the two duration points is: $\frac{dur_1}{dur_2}$, where dur_1 and dur_2 are respectively the durations of the syllable as pronounced by the L1 speaker and by the L2 speaker.

The f_0 value of the Momel anchor points are similarly modified on the pitch tier to $f0_1 * \frac{median_2}{median_1}$), where $f0_1$ is the f_0 value of the anchor point for the L1 speaker and $median_1$ and $median_2$ are the f_0 medians for the two speakers. These manipulations modify the L2 utterance so that the durations correspond to those of the L1 utterance and the f_0 contour corresponds to that of the L1 speaker, shifted logarithmically so that it is centred on the L2 speaker's median f_0. For best results these anchor points are then interpolated quadratically.

The time values of the anchor points are those of the L1 speaker. These do not need to be modified since the duration of the L2 speaker's syllables have been adjusted to be the same as that of the L1 speaker.

The output of this modification, resynthesised by Praat with the overlap and add algorithm, would allow the learner to hear her/his own voice with the prosody of the native speaker which was prosodically most similar to that of the L2 recording.

7. Conclusions and perspectives

All of the steps described above can now be carried out fully automatically with freely available tools. The tool is necessarily limited to read speech, because of the need for pre-recorded models. It seems likely, though, that this type of feedback could be extremely beneficial to language learners at all levels.

We hope in future work to apply this technique systematically to larger samples of L2 speech for both Chinese speakers and for other languages. We intend in particular to look into the possibility of quantifying the automatic evaluation of an L2 speaker's prosody and looking at how well this automatic evaluation corresponds to the subjective evaluation by language teachers. We intend furthermore to investigate the effectiveness of a tool such as that described in the previous sections, providing visual and auditory feedback to a language learner.

Figure 4: Prosody of the utterance T02-S3 "*Could you arrange to send an engineer on Tuesday morning please*?" by one (female) native English speaker.

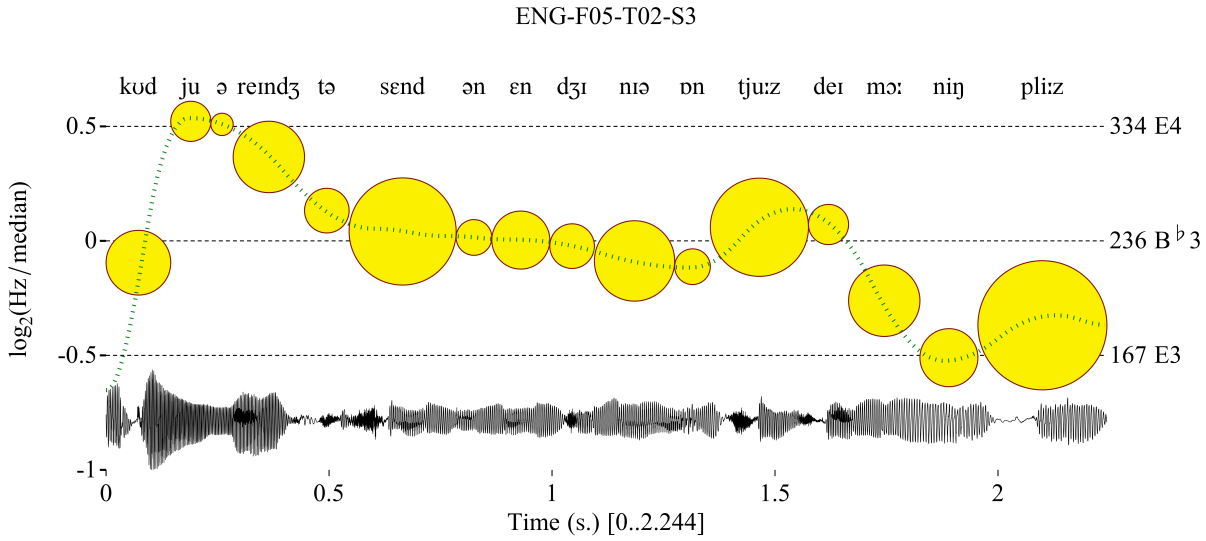
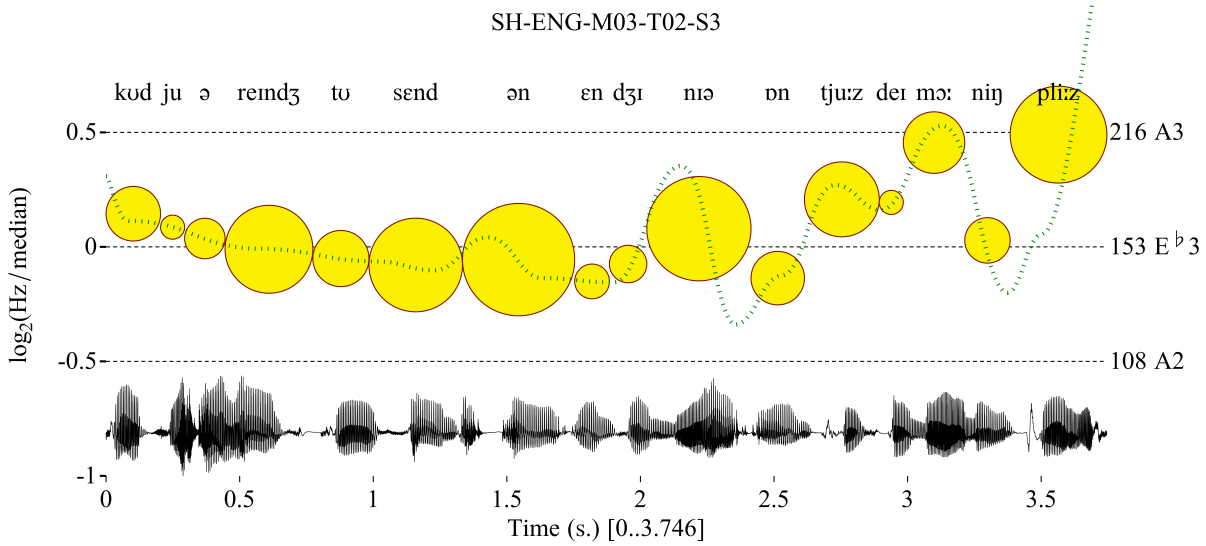


Figure 5: Prosody of the utterance T02-S3 "*Could you arrange to send an engineer on Tuesday morning please*?" by one (male) Chinese speaker.



8. Acknowledgements

My grateful thanks to Professor Qiuwu Ma from Tongji University and Professor Hongwei Ding, now at Shanghai Jiao Tong University for their support during my three year contract (2011-2014) as Lecture Professor at the School of Foreign Languages, Tongji University, Shanghai, as well as to Xinping Xu from Tongji University, for her help in the collection of the data for Shanghai speakers.

9. References

[1] M. Jilka 2000. *The contribution of intonation to the perception of foreign accent.* Ph.D. dissertation, Institut fr Maschinelle Sprachverarbeitung, University of Stuttgart.

[2] Hahn, L. 2004. Primary stress and intelligibility: research to motivate teaching of suprasegmentals. *TESOL Quarterly*, 38, 201 223.

[3] Hirst, D.J. 2013. Melody metrics for for prosodic typology: comparing English, French and Chinese. in *Proceedings of Interspeech (International Conference on Speech Processing).*, Lyon, August 2013.

[4] Hirst, D.J.; Ding, H.. 2015. Using melody metrics to compare English speech read by native speakers and by L2 Chinese speakers from Shanghai. In *Proceedings of Interspeech (International Conference on Speech Processing)*, Dresden, September 2015.

[5] Chan, D. Fourcin, A.; Gibbon, D.; Granstrom, B.; Huckvale, M.; Kokkinakis, G.; Kvale, K.; Lamel, L.; Lindberg, B.; Moreno, A.; Mouropoulos, J.; Senia, F.; Trancoso, I.; Veld, C.; Zeiliger, J. 1995. Eurom - a spoken language resource for the EU. In *Proceedings Eurospeech 4,*, 1, 867870, Madrid, 18-21 September

[6] Kim, S-H.; Hirst, D.J.; Cho, H.-S.; Lee, H.-Y. and Chung, M.-H. 2008. Korean Multext: A Korean prosody corpus. In *Proceedings of the 4th International Conference on Speech Prosody*, Campinas, Brazil.,

[7] Herment, S.; Tortel, A.; Bigi, B.; Hirst, D.J.; Loukina, A. 2012. AixOx: A multi-layered learners corpus: automatic annotation. *4th International Conference on CorpusLinguistics.*, Jan, Spain, (in Daz Prez, J. and Daz Negrillo, A. (eds.) 2014. *Specialisation and variation in language corpora*, Peter Lang.)

[8] Hirst, D.J.; Bigi, B.; Cho, H.-S.; Ding, H.; Herment, S.; Wang, T. 2013. Building OMProDat, an open multilingual prosodic database. in *Proceedings of TRASP, Tools and Resources for the Analysis of Speech Prosody* [satellite workshop of Interspeech] Aix-en- Provence, August 2013, pp. 1114. http://sldr.org/sldr000725

[9] Hirst, D.J. 2007. A Praat Plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation., in *Proceeedings ICPhS* 2007. 1233-1236.

[10] Bigi, B.; Hirst, D. J. 2012. SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In *Proceedings of the 6th International Conference on Speech Prosody.*, Shanghai, May 2012.

[11] Pulgram, E. 1970. *Syllable, Word, Nexus, Cursus*. The Hague : Mouton.

[12] De Looze, C.; Hirst, D.J. 2014. The OMe (Octave-Median) scale: a natural scale for speech prosody. in *Proceedings of the 7th International Conference on Speech Prosody (SP7)*, N. Campbell; D. Gibbon; D. J. Hirst, Eds., Trinity College, Dublin, Ireland, May 2014.

[13] Hirst, D.J. 2015. ProZed: A Speech Prosody Editor for Linguists, Using Analysis-by-Synthesis in Hirose, K.; Tao, J. (eds) 2015. *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis*. Springer Verlag, Berlin Heidelberg. pp 3-17

[14] OConnor, J.D.; Arnold, G. 1961. *Intonation of Colloquial English. A Practical Handbook.* Longmans, London,

[15] 2006. Wells, John *English Intonation: An Introduction*. Cambridge University Press, Cambridge

[16] Boersma, P.; Weenink, D. 1992 (2015). *Praat, a system for doing phonetics by computer*. http://www.praat.org [version 6.0.05, November 2015]